2020 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21–24, 2020, ESPOO, FINLAND

DEEP CONVOLUTIONAL NEURAL NETWORK-BASED INVERSE FILTERING APPROACH FOR SPEECH DE-REVERBERATION

Hanwook Chung[1,2]*, Vikrant Singh Tomar*[2] *and Benoit Champagne*[1]

[1] Dept. of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada
[2] Fluent.ai, Montreal, QC, Canada
email: hanwook.chung@fluent.ai, vikrant.tomar@fluent.ai, benoit.champagne@mcgill.ca

ABSTRACT

In this paper, we introduce a spectral-domain inverse filtering approach for single-channel speech de-reverberation using deep convolutional neural network (CNN). The main goal is to better handle realistic reverberant conditions where the room impulse response (RIR) filter is longer than the short-time Fourier transform (STFT) analysis window. To this end, we consider the convolutive transfer function (CTF) model for the reverberant speech signal. In the proposed framework, the CNN architecture is trained to directly estimate the inverse filter of the CTF model. Among various choices for the CNN structure, we consider the U-net which consists of a fully-convolutional auto-encoder network with skip-connections. Experimental results show that the proposed method provides better de-reverberation performance than the prevalent benchmark algorithms under various reverberation conditions.

Index Terms— single-channel speech de-reverberation, inverse filtering, convolutive transfer function, deep convolutional neural network, U-net

1. INTRODUCTION

When capturing speech from a talker in an enclosed space, a microphone receives multiple delayed and attenuated copies of the original speech signal, caused by the reflections from walls, ceiling and floors, etc. [1]. The general objective of speech de-reverberation algorithms is to remove such reflected components from a reverberant speech signal while preserving the direct-path component to improve its quality and intelligibility. Speech de-reverberation has been an attractive research area and finds various applications, including mobile telephony, hearing aid and automatic speech recognition. A considerable amount of research efforts has been devoted to this problem in the past decades, leading to various approaches, such as spectral subtraction [2, 3], linear prediction-based approahces [3]-[5] and Kalman filtering [6]. However, these methods were originally introduced by using minimal amount of *a priori* information about the acoustic environment, specified by the room impulse response (RIR) between the speech source and the microphone. Consequently, they tend to provide limited de-reverberation performance under adverse conditions, e.g., a high level of reverberation or time-varying RIR.

In recent years, deep learning (DL)-based algorithms with strong nonlinear modeling capabilities have attracted enormous interest [7]. They have found diverse applications such as image classification [8], automatic speech recognition [9], speech enhancement [10]-[12] and speech de-reverberation [13]-[20], where they have shown remarkable performance. In general, supervised DL aims at estimating the nonlinear mapping function that relates the input features to the target features. In [13], a fully-connected multi-layer perceptron (MLP) is trained to directly predict the clean speech magnitude spectrum from a noisy reverberant speech magnitude spectrum. This type of approach, which aims to uncover the nonlinear relationship between the input and target features, has been further extended using various deep neural network (DNN) architectures, e.g., long short-term memory (LSTM) units [14], convolutional neural network (CNN) and generative adversarial network (GAN) [15]. Instead of directly estimating the clean speech features, more robust masking-based approaches have been introduced, e.g., direct estimation of a complex-valued ideal ratio mask (IRM) via MLP [16], implicit estimation of a real-valued IRM based on the late reverberation power spectral density (PSD) obatined via MLP [17], and phase-sensitive mask via GAN [18]. Some references focus on a feature-aware training framework that utilizes additional features as input to the DNN, such as reverberation time [19] or the late reverberation PSD [20]. The above DNN-based de-reverberation algorithms are implemented in the spectral-domain based on the assumption that the RIR filter length is smaller than the short-time Fourier transform (STFT) analysis window. In a real world scenario, however, such an assumption is not valid and consequently, provide limited de-reverberation performance.

In this paper, to overcome the above limitation, we introduce a novel spectral-domain inverse filtering approach for single-channel speech de-reverberation using a deep CNN. The main goal is to better handle realistic reverberant conditions, i.e., where the RIR filter length exceeds the STFT window length. To this end, we consider the convolutive transfer function (CTF) model [21], where the reverberant speech spectrum is represented by convolving the clean speech spectral coefficients with spectral filter coefficients along the time frame dimension for each frequency bin. In the proposed framework, we train the CNN architecture to directly estimate the inverse filter of the CTF model. During the de-reverberation stage, the estimated inverse filter is applied to the reverberant speech spectrum to obtain the clean speech spectrum. Among various choices for the CNN structure, we use the U-net consists of a fully-convolutional auto-encoder network with skip-connections [22]. Specifically, motivated by [11], we consider an online U-net structure for estimating the inverse filter for each time frame to better handle the time-varying RIR conditions. Objective experimental results show that the proposed method provides better de-reverberation performance than the prevalent benchmark algorithms under various room reverberation conditions.

Funding for this work was provided by grants from NSERC and Mitacs (Canada), with sponsorship from Fluent.ai (Montreal, Canada).

978-1-7281-6662-9/20/$31.00

2. REVERBERANT SIGNAL MODEL

Let us denote by y_n the observed reverberant speech signal at the discrete-time index $n \in \{0, 1, ..., N-1\}$. In the single-channel speech de-reverberation problem, by taking into account the convolutive nature of the acoustic medium as represented by the RIR between the speech source and the microphone, the reverberant speech signal can be written in the time-domain as

$$y_n = \sum_{q=0}^{Q-1} h_{nq} s_{n-q}, \tag{1}$$

where h_{nq} is the time-varying RIR filter coefficient at time n and delay index $q \in \{0, ..., Q-1\}$ and s_n is the clean speech signal. Considering the propagation delay between the source and microphone, the reverberant speech signal can be divided into three components: direct-path, early reverberant and late reverberant signals. On this basis, the signal model in (1) can be rearranged as

$$y_n = \underbrace{\sum_{q=0}^{Q_e-1} h_{nq} s_{n-q}}_{\triangleq y_n^E} + \underbrace{\sum_{q=Q_e}^{Q-1} h_{nq} s_{n-q}}_{\triangleq y_n^L}, \tag{2}$$

where y_n^E is the sum of the direct-path and early reverberant signals (hereafter referred to as the early reverberant signal for simplicity), y_n^L is the late reverberant signal, and Q_e is the RIR filter length corresponding to early reverberation signal, i.e., the filter index separating the RIR into early and late reverberation components. It has been shown that the late reverberation components are the major cause of the degradation of the speech intelligibility [23]. In this paper, hence, we focus on reducing the late reverberation components, while aiming at recovering the early reverberant signal from the reverberant speech. Such an algorithm is commonly referred to as late reverberation suppression.

In audio and speech signal processing, the frequency-domain representation is commonly used in order to better exploit spectral characteristics, where a popular choice is the STFT. Numerous spectral-domain speech de-reverberation algorithms assume that the RIR filter length is much smaller than the STFT analysis window. In this case, the effect of reverberation in the frequency domain amounts to a simple multiplication of the room transfer function and the clean speech spectral coefficients. Such an assumption, however, is often not valid in a real world scenario and hence, may lead to limited de-reverberation performance. To overcome this limitation, we consider a more comprehensive CTF model [21]:

$$Y_{kl} = \sum_{p=0}^{P-1} H_{klp}^{*} S_{k,l-p}, \tag{3}$$

where $Y_{kl} \in \mathbb{C}$ and $S_{kl} \in \mathbb{C}$ respectively denote the STFT coefficients of the reverberant and clean speech signals at the frequency bin $k \in \{0, ..., K-1\}$ and time frame $l \in \{0, ..., L-1\}$, $H_{klp} \in \mathbb{C}$ is the time-varying CTF coefficient with frame delay index $p \in \{0, ..., P-1\}$, and the superscript $*$ denotes complex conjugation.

3. PROPOSED DNN-BASED DE-REVERBERATION

In the proposed framework, we aim to directly estimate the inverse filter of the CTF model based on an online U-net architecture. In this section, after explaining the proposed inverse filtering approach, we describe the online U-net structure and its application to de-reverberation.

3.1. Proposed inverse filtering approach

The clean speech spectral coefficients, S_{kl}, can be estimated via inverse filtering of the CTF model [24]:

$$\hat{S}_{kl} = \sum_{p=0}^{P_d-1} \widetilde{W}_{klp}^{*} Y_{k,l-p}, \tag{4}$$

where $\widetilde{W}_{klp} \in \mathbb{C}$ is the complex-valued time-varying inverse filter coefficient with frame delay index $p \in \{0, ..., P_d-1\}$. In this paper, we propose a novel inverse filtering method by applying two modifications to (4) as follows. First, instead of estimating the clean speech, we aim at estimating the early reverberant signal, since the latter is sufficient to improve the speech intelligibility as mentioned in Sec. 2. Furthermore, the suppression of late reverberation is not affected by te misalignment between the observed reverberant and the clean speech signal, which is caused due to the direct-path propagation delay between the source and microphone. Consequently, this provides a more robust de-reverberation performance. Second, instead of estimating complex-valued spectral coefficients, we focus on estimating the spectral magnitudes, as the latter components are knwon to contribute more towards speech intelligibility than the phase components [25][1]. In this way, we can also reduce the computational cost compared to handling both the magnitude and phase components in general. Hence, taking into account the above modifications, we propose the following inverse filtering model to estimate the magnitude spectral coefficients of the early reverberant signal:

$$|\hat{Y}_{kl}^E| = \sum_{p=0}^{P_d-1} W_{klp} |Y_{k,l-p}|, \tag{5}$$

where $W_{klp} \in \mathbb{R}$ is the *real-valued* time-varying inverse filter.

3.2. Online U-net architecture for inverse filtering

In the proposed framework, we estimate the inverse filter in (5) using a CNN structure. The CNN transforms given input features through a series of hidden layers, based on the convolution operation. Let $i = \{1, ..., I\}$ denote the hidden layer index, while $i = 0$ and $i = I+1$ indicate the input and output features, respectively. The i-th hidden layer output is computed by convolving the $(i-1)$-th hidden layer output with a filter, also referred to as kernel, followed by a non-linear transformation via an activation function. The convolution operation enables to extract local patterns of the given features efficiently, as observed in adjacent time-frequency bins in the STFT domain. Moreover, the CNN architecture generally requires less parameters, i.e., the number of kernel coefficients, compared to a fully-connected MLP and hence, is known to better handle the over-fitting problem. The resulting hidden layer output can be represented by a $C_i \times K_i \times L_i$ tensor, where C_i, K_i and L_i respectively denote the number of channels, frequency bins and time frames for the i-th hidden layer.

Among various choices for the CNN structure, we use the U-net which consists of a fully-convolutional auto-encoder (CAE) with skip-connections [22]. The CAE consists of two stages: an *encoder*, which compresses the given features into a lower dimensional space by capturing their key attributes; and a *decoder*, which expands the compressed features (also known as bottleneck) features into a desired feature space. The skip-connection method uses the i-th hidden

[1] As recent studies taking the phase components into account have shown promising results, e.g., speech enhancement using complex U-net architecture [26], such an approach remains an interesting avenue for our future work.

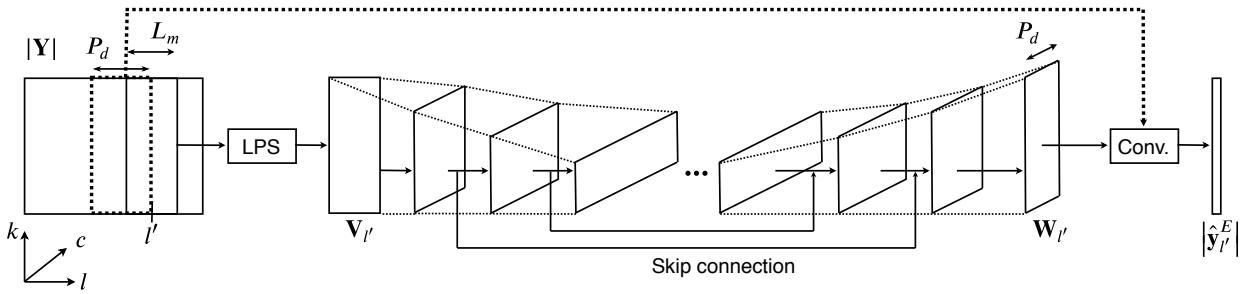

Fig. 1. The architecture of the online U-net for the proposed inverse filtering-based speech de-reverberation.

layer output as an additional input feature for the $(I-i)$-th hidden layer. The main advantage of using skip-connection is that it can handle the vanishing gradient issue, which results in an ineffective update of the lower hidden layer parameters due to an extremely small gradient value while implementing error back-propagation.

In the propose framework, motivated by [11], we consider an online U-net structure for estimating an inverse filter for each time to better handle the time-varying RIR conditions. Regarding the input features, we use the log power spectral coefficients (LPS) of the reverberant speech, i.e., $v_{kl} \triangleq \ln(|Y_{kl}|^2)$. In order to better handle the temporal dependencies, we construct a multi-frame input feature matrix, $\mathbf{V}_l \in \mathbb{R}^{K\times L_m}$, by concatenating feature vectors from L_m successive time frames (e.g., [11]) as follows:

$$\mathbf{V}_l = [\mathbf{v}_{l-(L_m-1)/2}, .., \mathbf{v}_l, ..., \mathbf{v}_{l+(L_m-1)/2}], \tag{6}$$

where $\mathbf{v}_l = [v_{kl}] \in \mathbb{R}^K$ and we consider an odd number for L_m. The output of the online U-net is the inverse filter for the given l-th frame, i.e., $\mathbf{W}_l = [W_{klp}] \in \mathbb{R}^{K\times P_d}$. Specifically, we align the frame delay dimension of the inverse filter coefficients along the CNN channel axis for a practical implementation, i.e., $C_{I+1} = P_d$ (a more detailed explanation will be presented in Sec. 4.2). The online U-net structure for the proposed inverse filtering-based de-reverberation is illustrated in Fig. 1. Note that the tensor form features are expressed in a 3-D $k \times l \times c$ coordinate system, where k, l and c respectively indicate the frequency bin, time frame and channel index.

During the proposed training stage, the online U-net parameters, i.e., the kernel coefficients, are estimated by minimizing the mean-square error (MSE):

$$E = \frac{1}{KL}\sum_{k=0}^{K-1}\sum_{l=0}^{L-1}\left(|Y^E_{kl}| - |\hat{Y}^E_{kl}|\right)^2, \tag{7}$$

where $|\hat{Y}^E_{kl}|$ is computed using (5). During the proposed de-reverberation stage, we estimate the complex-valued early reverberant spectrum by combining the magnitude components estimated via (5) and the phase components from the reverberant speech, i.e.,

$$\hat{Y}^E_{kl} = \left(\sum_{p=0}^{P_d-1}\hat{W}_{klp}|Y_{k,l-p}|\right)e^{j\angle Y_{kl}}, \tag{8}$$

where $j = \sqrt{-1}$. Finally, the de-reverberated speech signal in the time-domain is reconstructed by applying the inverse STFT to $\hat{Y}^E_{kl}$, followed by the overlap-add method.

4. EXPERIMENTS

In this section, after describing the data sets and general methodology, we present and discuss the experimental results.

4.1. Data sets

We conducted experiments using the clean speech from the TIMIT corpus [27], where the sampling rate of all signals was 16 kHz. Regarding the RIR, we employed two data sets: *simulated* RIRs via the RIR generator [28], and *real-measured* RIRs from the C4DM database [29]. The former are obtained based on the image method for a given 3-D rectangular room, reverberation time RT_{60}, and source-microphone positions. The latter are collected from GreatHall, Octagon and Classroom using the logarithmic sine sweep method, where the measured reverberation time RT_{30} of all rooms was approximately 2s (see [29] for more details). The speech and RIR files were divided into three *disjoint* groups: i) *training data*, used for estimating the U-net parameters; ii) *validation data*, used for selecting tuning parameters such as the multi-frame length L_m and inverse filter length P_d; and iii) *test data*, used during the de-reverberation stage to evaluate the performance. For all data, the reverberant speech signals were obtained by convolving the clean speech signals with the RIRs.

Regarding the training data, we selected 4620 utterances from the "train" set as the clean speech. For the simulated RIRs, we considered a room with size of $8 \times 6 \times 4$ m (measured along Cartesian coordinate axes), which will be referred to as Room 1, and reverberation times RT_{60} of 500, 750 and 1000 ms. We generated 15 RIRs for each reverberation time by varying the source-microphone positions, resulting in a total of 45 RIRs. For the real-measured RIRs, we selected 50 RIRs from GreatHall, 50 RIRs from Octagon and 40 RIRs Classroom, resulting in a total of 140 RIRs. Regarding the validation data, we selected 400 utterances from the "test" set as the clean speech. We generated 10 RIRs from Room 1 with RT_{60} of 500, 750 and 1000 ms. We selected 5 RIRs from each one of GreatHall, Octagon and Classroom.

Regarding the clean speech test data, we selected 192 utterances, from the "test" set. For the simulated RIRs, we generated 10 RIRs from Room 1 with RT_{60} of 500, 750 and 1000 ms. To evaluate the performance for an unseen type of acoustic environment, we additionally generated 10 RIRs from a room with size of $6 \times 4 \times 3.5$ m, which will be referred to as Room 2, and reverberation times RT_{60} of 500, 750 and 1000 ms. For the real-measured RIRs, we selected 10 RIRs from each one of GreatHall, Octagon and Classroom. Besides the above *static* RIR conditions, we additionally considered *time-varying* RIR scenarios. To this end, we divided 10 RIRs into

Table 1. Average results for the static simulated RIRs

Room type	RT_{60} (ms)	Eval.	Rev.	DSM	iIRM	dIRM	iFilt
Room 1	500	SDR	2.73	2.73	3.86	1.35	**4.19**
		ESTOI	0.49	0.62	0.52	0.58	**0.67**
		SRMR	2.69	2.98	2.63	2.41	**3.59**
	750	SDR	-0.56	1.11	1.27	0.14	**2.25**
		ESTOI	0.33	0.52	0.38	0.50	**0.55**
		SRMR	2.14	2.61	2.23	2.28	**3.24**
	1000	SDR	-2.52	-0.20	-0.59	-0.52	**0.93**
		ESTOI	0.26	0.46	0.30	0.45	**0.48**
		SRMR	1.79	2.39	1.97	2.24	**3.01**
Room 2	500	SDR	3.17	2.81	4.02	1.07	**4.19**
		ESTOI	0.49	0.63	0.53	0.57	**0.67**
		SRMR	2.77	3.03	2.72	2.41	**3.66**
	750	SDR	0.58	1.83	2.45	0.52	**3.12**
		ESTOI	0.36	0.55	0.41	0.53	**0.59**
		SRMR	2.21	2.81	2.40	2.36	**3.45**
	1000	SDR	-1.64	0.43	0.48	-0.27	**1.58**
		ESTOI	0.28	0.49	0.32	0.48	**0.51**
		SRMR	1.83	2.56	2.08	2.29	**3.18**

Table 2. Average results for the time-varying simulated RIRs

Room type	RT_{60} (ms)	Eval.	Rev.	DSM	iIRM	dIRM	iFilt
Room 1	500	SDR	-0.84	0.00	0.42	-0.34	**0.73**
		ESTOI	0.48	0.62	0.52	0.58	**0.66**
		SRMR	2.71	2.96	2.67	2.48	**3.61**
	750	SDR	-3.30	-1.01	-1.41	-1.04	**-0.12**
		ESTOI	0.34	0.52	0.38	0.50	**0.55**
		SRMR	2.10	2.61	2.18	2.31	**3.23**
	1000	SDR	-5.09	-2.72	-3.28	-2.62	**-1.90**
		ESTOI	0.24	0.45	0.28	0.45	**0.47**
		SRMR	1.74	2.37	1.94	2.21	**2.97**
Room 2	500	SDR	-0.61	-0.46	**0.16**	-1.71	0.12
		ESTOI	0.49	0.63	0.53	0.57	**0.67**
		SRMR	2.76	2.97	2.69	2.37	**3.58**
	750	SDR	-2.92	-1.57	-1.30	-2.39	**-0.90**
		ESTOI	0.32	0.53	0.38	0.51	**0.56**
		SRMR	2.11	2.77	2.28	2.33	**3.35**
	1000	SDR	-3.96	-1.79	-1.94	-2.07	**-1.01**
		ESTOI	0.26	0.47	0.31	0.46	**0.49**
		SRMR	1.72	2.43	1.98	2.21	**3.03**

two groups, allowing us to generate two different time-varying RIRs scenarios, each comprised of 5 RIRs. Specifically, for each scenario, the reverberant speech was obtained by convolving the given speech utterance with one of the 5 RIRs in cycle for every 1s.

4.2. Methodology

Regarding the implementation of the proposed de-reverberation algorithm, a Hamming window of 400 samples with 60% overlap and 512-point fast Fourier transform (FFT) were employed for the STFT analysis. We set the multi-frame length to $L_m = 5$, the inverse filter length to $P_d = 9$, and the early reverberation filter lengths to $Q_e \in \{32, 64, 128\}$. We designed the U-net using $I = 11$ hidden layers with corresponding numbers of channels $[16, 16, 32, 32, 64, 64, 64, 32, 32, 16, 16]$ and $C_{I+1} = P_d = 9$ for the output layer, based on the validation data. We used the rectified linear units (ReLU) as the activation function for all hidden layers and linear activation function for the output layer. Batch normalization was applied to all hidden layers [30]. The U-net parameters were updated iteratively via error back-propagation and the adaptive moment estimation (Adam) optimizer [31], with the mini-batch size of 32 for a total of 200 epochs. The initial learning rate was set to 0.001, which decreased by 10% for every 10 epochs. We used the kernel size of 9 with stride of 2 along the k-axis for all hidden and output layers. To ensure the non-negativity of the estimated spectral magnitudes of the early reverberant speech, i.e., $|\hat{Y}_{kl}^E| \geq 0$, we applied ReLU to the inverse-filtered reverberant speech in (5). In order to efficiently implement the proposed online U-net for inverse filtering, we considered the 2-D convolution operation for all layers as follows. For the given input feature matrix $\mathbf{V} \in \mathbb{R}^{K \times L}$, we applied the kernels of sizes $9 \times L_m$ to the input layer $i = 0$, and 9×1 to all hidden layers $i \in \{1, ..., I\}$, with stride of 2 along the k-axis and 1 along the l-axis. The number of time frames was set to $L_i = L$ for all $i \in \{0, 1, ..., I\}$, and the size of the output layer was $C_{I+1} \times K_{I+1} \times L_{I+1} = P_d \times K \times L$. The time-shifted magnitude spectra of the reverberant speech were concatenated along the c-axis to obtain a tensor of size $C_0 \times K_0 \times L_0 = P_d \times K \times L$. The estimated early reverberant magnitude spectrum given by (5) was then computed by multiplying the above U-net input and output tensors, followed by summing along the c-axis.

To evaluate the performance of the proposed method, we implemented several benchmark algorithms: i) direct estimation of the clean speech magnitude spectral coefficients (DSM) [15], ii) implicit estimation of a real-valued IRM based on the late reverberation PSD obtained via DNN (iIRM) [17], and iii) direct estimation of a real-valued IRM (dIRM) [32]. Specifically, we considered the LPS of the early reverberant signal as the target output feature for DSM. The target IRM in the dIRM method was constructed based on the well-known Wiener filter, specified by the PSDs of the early and late reverberant signals, i.e., $|Y_{kl}^E|^2/(|Y_{kl}^E|^2 + |Y_{kl}^L|^2)$. Although the iIRM and dIRM methods were originally proposed using a fully-connected MLP, we implemented them using the online U-net as explained in Sec. 3.2 for fair comparison. Basic settings such as the STFT analysis and synthesis, the U-net configuration, the input feature type (i.e., the LPS of the reverberant speech) and the mini-batch size were kept identical when applicable.

To evaluate the de-reverberation performance, we considered the source-to-distortion ratio (SDR) [33], extended short-time objective intelligibility (ESTOI) [34] and speech-to-reverberation modulation energy ratio (SRMR) [35] as the objective measures. The SDR is computed in dB based on the source-to-interference ratio (SIR) and source-to-artifact ratio (SAR), and has been widely used in audio source separation and speech enhancement, e.g., [11, 12]. For a given target source signal, in general, the interference refers to unwanted signal such as late reverberation components, whereas the artifact refers to forbidden distortion. In speech de-reverberation applications, these measures can be interpreted as follows: the SIR and SAR are proportional to the amount of late reverberation suppression and inversely proportional to the clean speech distortion, respectively, while the SDR measure the overall quality of the de-reverberated speech signal. The ESTOI is computed based on the spectral correlation between the short-time auditory filter-bank coefficients of the target clean speech and processed speech, and has shown to be closely related to speech intelligibility of a human listener. The SRMR is a non-intrusive metric for speech quality and intelligibility based on an auditory-inspired modulation spectral representation of the speech signal. For all measures, a higher value indicates a better result.

Table 3. Average results for the static real-measured RIRs

Room type	Eval.	Rev.	DSM	iIRM	dIRM	iFilt
GreatHall	SDR	-2.36	-0.59	0.27	1.25	**1.84**
	ESTOI	0.28	0.39	0.30	0.42	**0.45**
	SRMR	1.38	2.36	2.14	2.50	**3.32**
Octagon	SDR	-2.95	-0.84	0.64	1.02	**1.62**
	ESTOI	0.29	0.39	0.30	0.42	**0.46**
	SRMR	1.28	2.19	2.21	2.26	**3.06**
Classroom	SDR	-4.08	-1.93	-1.96	-1.52	**-0.16**
	ESTOI	0.19	0.36	0.23	0.36	**0.40**
	SRMR	1.16	2.12	1.81	2.00	**2.86**

Table 4. Average results for the time-varying real-measured RIRs

Room type	Eval.	Rev.	DSM	iIRM	dIRM	iFilt
GreatHall	SDR	-5.42	-3.02	-2.06	-1.57	**-1.37**
	ESTOI	0.23	0.36	0.26	0.38	**0.42**
	SRMR	1.37	2.41	2.12	2.41	**3.33**
Octagon	SDR	-4.59	-2.63	-1.81	-0.70	**-0.45**
	ESTOI	0.26	0.37	0.28	0.40	**0.44**
	SRMR	1.14	2.05	2.00	2.20	**2.91**
Classroom	SDR	-6.07	-3.98	-4.19	-3.87	**-2.54**
	ESTOI	0.19	0.36	0.22	0.36	**0.40**
	SRMR	1.14	2.10	1.39	1.99	**2.84**

4.3. Results

The average results using the static and time-varying simulated RIR, in case of an early reverberation RIR filter length of $Q_e = 32$, are shown in Tables 1 and 2, respectively. The proposed inverse filtering-based approach is referred to as iFilt. The values in bold indicate the best performance along the corresponding row. As we can see, the propose method provided better de-reverberation performance than the benchmark algorithms for all room types and both the static and time-varying RIR conditions, in general. The only exception was found from the SDR value for the time-varying RIR in Room 2 at $RT_{60} = 500$ ms, where the iIRM method provided slightly better result.

The average results using the static and time-varying real-measured RIRs from the C4DM database, in case of an early reverberation RIR filter length of $Q_e = 32$, are shown in Tables 3 and 4, respectively. As we can see, the proposed method provided the best results for all room types as well as for static and time-varying RIRs. Interestingly, the dIRM method provided better results than the DSM and iIRM methods in general especially in terms of the SDR value, in contrast to the results found when using the simulated RIRs in Tables 1 and 2.

In the following, we comment on some additional experimental results, which we did not report in this paper due to space limitation. First, we observed that the proposed method provided better performance than the benchmark algorithms for $Q_e = 64$ and 128, following similar trend to those reported above for $Q_e = 32$. Second, we were able to verify that estimating the early reverberant signal y_n^E resulted in better de-reverberation performance than attempting to estimate the clean speech signal s_n.

5. CONCLUSION AND FUTURE WORKS

We introduced a spectral-domain inverse filtering approach for single-channel speech de-reverberation using a DNN. The main goal was to better handle realistic reverberant conditions where the RIR filter is longer than the STFT analysis window. To this end, we considered the CTF model for the reverberant speech signal. In the proposed framework, we aimed at estimating the magnitude spectral coefficients of the early reverberant speech signal via inverse filtering of the CTF model. The inverse filter was estimated based on the online U-net architecture, which consists of a fully-convolutional CAE with skip-connections. We conducted experiments using both the simulated and real-measured RIRs. Experiments showed that the proposed method provides better de-reverberation performance than the prevalent benchmark algorithms under various reverberation conditions, i.e., different levels of reverberation time, unseen type of room environment, static and time-varying RIR conditions, and for the simulated and real-measured RIRs.

Several avenues remain opened for future research. First, we can extend the proposed method to a complex-valued U-net architecture to handle the phase components [26]. Second, we can incorporate additional information, such as reverberation time [19] or late reverberation PSD [20], into the proposed inverse filtering framework. Besides the above avenues, which are mainly for further improving the speech quality of the reverberant speech, it would also be of interest to evaluate experimentally the effects of the proposed de-reverberation algorithm when used as a front-end for automatic speech recognition.

6. REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, Springer Science & Business Media, 2010.

[2] K. Lebart, J. M. Boucher and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acustica*, vol. 87, no. 3, pp. 359366, May 2001.

[3] M. Wu and D. L. Wang, "A two-stage algorithm for one microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 3, pp. 774784, Apr. 2006.

[4] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi and B. -H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, pp. 1717-1731, Sep. 2010.

[5] M. Parchami, W.-P. Zhu and B. Champagne, "Speech dereverberation using linear prediction with estimation of early speech spectral variance," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 504-508, Mar. 2016.

[6] B. Schwartz, S. Gannot and E. A. P. Habets, "Online speech dereverberation using Kalman filter and EM algorithm," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 23, no. 2, pp. 394-406, Feb. 2015.

[7] Y. Bengio, A. Courville and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no 8, pp. 1798-1828, Aug. 2013.

[8] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 770-778, June 2016.

[9] L. Deng, G. Hinton and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Proc. IEEE Int. Conf. Acoustics,*

Speech, and Signal Process. (ICASSP), pp. 8599-8603, May 2013.

[10] Y. Xu, J. Du, L. -R. Dai and C. -H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 23, no. 1, pp. 7-19, Jan. 2015.

[11] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. Interspeech*, pp. 1993-1997, Aug. 2017.

[12] H. Chung, T. Kim, E. Plourde and B. Champagne, "Noise-adaptive deep neural network for single-channel speech enhancement," in *Proc. Int. Workshop on Machine Learning for Signal Process. (MLSP)*, six pages, Sep. 2018.

[13] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 23, no. 6, pp. 982-992, June 2015.

[14] Y. Zhao, D. Wang, B. Xu and T. Zhang, "Late reverberation suppression using recurrent neural networks with long short-term memory," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 5434-5438, Apr. 2018.

[15] O. Ernst, E. Chazan, S. Gannot and J. Goldberger, "Speech dereverberation using fully convolutional networks," in *Proc. Europian Signal Process. Conf. (EUSIPCO)*, pp. 390-393, Sep. 2018.

[16] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 25, no. 7, pp. 1492-1501, July 2017.

[17] I. Kodrasi and H. Bourlard, "Single-channel late reverberation power spectral density estimation using denoising autoencoders," in *Proc. Interspeech*, pp. 1319-1323, Sep. 2018.

[18] C. Li, T. Wang, S. Xu and B. Xu, "Single-channel speech dereverberation via generative adversarial training," *arXiv: 1806.09325*, June 2018.

[19] B. Wu, K. Li, M. Yang and C. -H. Lee, "Dereverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 25, no. 1, pp. 102-111, Jan. 2017.

[20] Y. Qi, F Yang and J. Yang, "A late reverberation power spectral density aware approach to speech dereverberation based on deep neural networks," in *Proc. Asia-Pacific Signal and Information Process. Association Annual Summit and Conf. (APSIPA-ASC*, pp. 1700-1703, Nov. 2019.

[21] R. Talmon, I. Cohen and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no.7, pp. 1420-1434, Sep. 2009.

[22] O. Ronnenberger, P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Medical Image Computing and Computer-assisted Intervention*, pp. 234-241, Oct. 2015.

[23] A. K. Nábělek, T. R. Letowski and F. M. Tucker, "Reverberant overlap-and self-masking consonant identification," *The Journal of the Acoustical Society of America*, vol. 86, no. 4, pp. 1259-1265, June 1989.

[24] X. Li, L. Girin, S. Gannot and R. Horaud, "Multichannel speech separation and enhancement using the convolutive transfer function," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 27, no. 3, pp. 645-659, Jan. 2019.

[25] D. O'Shaughnessy, *Speech Communications: Human and Machine*, IEEE Press, 1987.

[26] H. -S Choi, J. -H Kim, J, Huh, A. Kim, J. -W. Ha and K. Lee, "Phase-aware speech enhancement with deep complex U-net," in *Proc. Int. Conf. Learning and Representation (ICLR)*, 20 pages, May 2019.

[27] J. S. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren and V. Zue, *TIMIT: acoustic-phonetic continuous speech corpus*, Linguistic Data Consortium, 1993.

[28] E. A. P. Habets, *Room Impulse Response Generator*, Technische Universiteit Eindhoven, Tech. Rep., 2006.

[29] R. Stewart and M. Sander, "Database of omnidirectional and B-format impulse responses," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 165-168, Mar. 2010.

[30] S. Loffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv: 1502.03167. 2015*, Feb. 2015.

[31] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," *arXiv: 1412.6980*, Dec. 2014.

[32] Y. Wang, A. Narayanan and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 22, no. 12, pp. 18491858, Dec. 2014.

[33] E. Vincenet, R. Gribonval and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp. 1462-1469, July 2006.

[34] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 24, no. 11, pp. 2009-2022, Nov. 2016

[35] J. F. Santos, M. Senoussaoui and T. H. Falk, "An updated objective intelligibility estimation metric for normal hearing listeners under noise and reverberation," in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 55-59, Sep. 2014.